Report of Stat 5044 take-home exam

Author: Miao Yuan

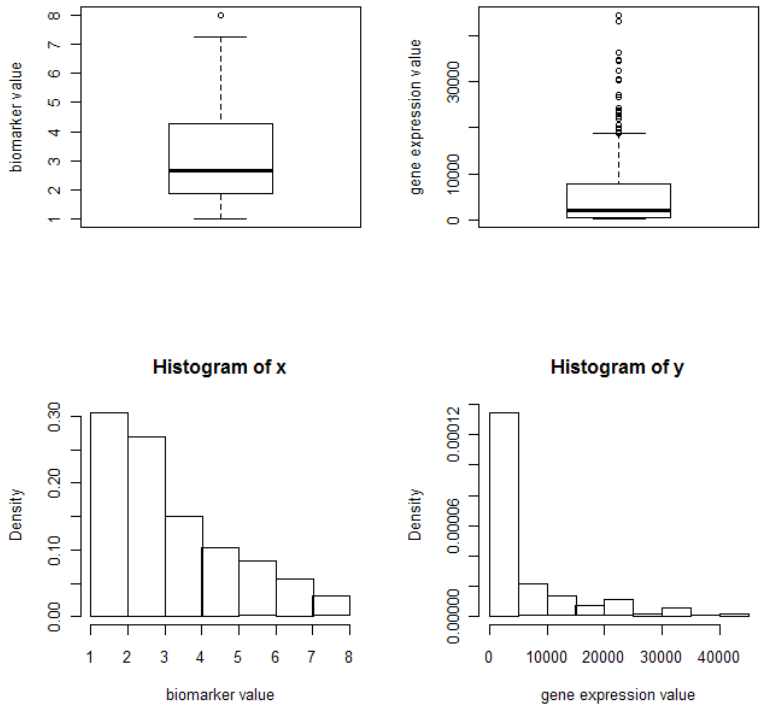
1. Introduction

In this problem, there are two variables: biomarker value of diabetes and a certain gene expression value. Since the gene expression values are more difficult to obtain in Biological experiments, I choose the biomarker value as explanatory variable and certain gene expression value as response variable. Thus, it is possible to use the biomarker value to predict the corresponding gene expression value through fitted model between the two variables.

1.1 Overview of data

Summary Statistics	minumum	25% quantile	median	mean	75% quantile	maximum	standard deviation
X(biomarker value)	1.000	1.890	2.660	3.189	4.260	8.000	1.717392
Y(gene expression value)	90	467	1938	6408	7850	44580	9377.946

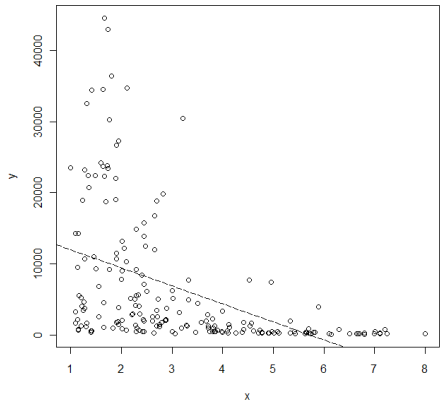
Table 1.1.1 Summary Statistics of the two variables



Graph1.1.1 Boxplots and histograms of the two variables

From the histogram of biomarker values, I see their distribution is significantly right-skewed; the fact that mean of biomarker values is larger than the median also implies the data is right-skewed. While, the boxplot of biomarker values indicate that there is one outlier in the data.

From the histogram of gene expression values, I see their distribution is highly right-skewed; the fact that mean of gene expression values is significantly larger than the median also implies the distribution has longer right tail. Besides, the gene expression values have large variance; the big range of data and the boxplot both indicate there may be some outliers in the data.



Graph 1.2 Scatter plot and simple linear function of the two variables

From graph1.2, I hardly see any linear pattern between biomarker values and gene expression values. The simple linear model does not seem to be a good fit. The calculated R-square is 0.2147, which means the proportion of reduced error by fitting regression line instead of just using $\bar{x}$ is only 21.47%; and MSE(which is the same as $\hat{\sigma}^2$) is 69422224, which shows the mean square of error is very large. From the two indexes I can draw the conclusion that simple linear regression is not a good fit of relationship between biomarker values and gene expression values.

1.2 Goal of data analysis

Based on the analysis of raw data, I see the long right-tail of the distributions of biomarker values and gene expression values, also the large range and many outliers of the distribution of gene expression values. Besides, there is weak linear association between the two variables.

Thus, the goal of data analysis is to fit the best models that can reasonably describe the relationship between two variables and also satisfy model assumptions, using statistical techniques. Then, the fitted models should be applied to do some inference, including estimating, calculating confidence intervals, testing parameters of fitted model and obtaining the point estimation, confidence interval and predict interval of the response variable given a new value of explanatory variable.

2. Methods to find the best models, Results and Conclusions

There exist some outliers in this model based on the results of cook-distance analysis. Outliers are significant factors in determining the degree of fitness of models; we should properly deal with them to obtain the best models. There is a dilemma in manipulating outliers: on one hand, they may not be mistakes and thus are precious resources to analyze the data; on the other hand, they may be results of some mistakes, such as miscalculation or mis-measurement, thus will lead to less fitted model. Since there is not enough evidence to decide which one is the case in this problem, I find the best model in either of the two situations.

2.1 Fit the best model without removing outliers.

2.1.1 Fit polynomial models

Since the simple linear regression is not a good fit of the relationship between biomarker values and gene expression values, I fit polynomial models from the quadratic model to the 11^{th} order model. The 11^{th} model is the one with the most R-square and the least MSE: R-square= 0.3971; MSE=$\hat{\sigma}^2$ =56250000.

Since the 11^{th} order polynomial model is too complex to be applicable, it also violates the constant-variance and normality assumptions as the lower order models, I do not pick it but to choose the 5^{th} order polynomial model. Then the assumptions such as constant-variance, randomness and normality assumptions should be checked.

Call: lm(formula = y ~ x + I(x^2) + I(x^3) + I(x^4) + I(x^5))

Residuals :Min	1Q	Median	3Q	Max
-12420.9	-4561.8	-859.1	1666.0	31165.0
Coefficients:	Estimate	Std Error	t value	Pr(>|t|)
(Intercept)	-56858.43	22277.17	-2.552	0.01150 *
x	117028.72	36269.58	3.227	0.00148 **
I(x^2)	-68778.47	21215.61	-3.242	0.00141 **
I(x^3)	17423.50	5641.07	3.089	0.00232 **
I(x^4)	-2008.38	692.45	-2.900	0.00417 **
I(x^5)	86.42	31.84	2.714	0.00727 **
Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'
Residual standard error: 7986 on 187 degrees of freedom				
Multiple R-squared: 0.2937		Adjusted R-squared: 0.2748		
F-statistic: 15.55 on 5 and 187 DF			p-value: 8.713e-13	

Table2.1.1 The fitted 5^{th} order polynomial model

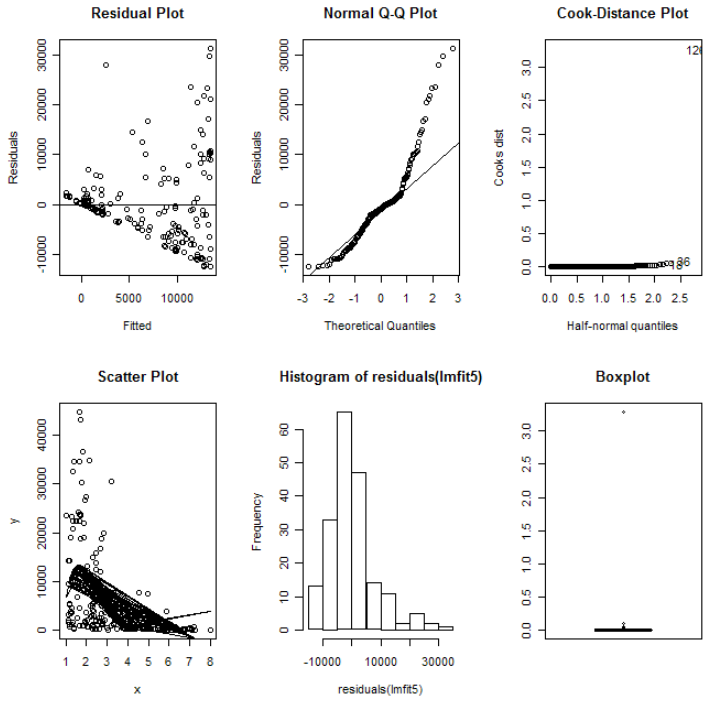
2.1.2 Check Assumptions

For the fitted 5^{th} order polynomial model, the p-values of run-test and Durbin-Watson test are 0.09685 and 0.6104, respectively, both of which are larger than 0.05, so the randomness assumption can be accepted at the significance lever 0.05.

The p-value of studentized Breusch-Pagan test is 8.081e-06, which is significantly smaller than 0.05 and the proportion that two random selected groups of residuals with the same sample size having different variances is 0.052, which is significantly larger than 0. Thus, we can reach the

conclusion that constant-variance assumption is violated.

The p-value of Shapiro-Wilk normality test is 8.288e-11, which is far less than 0.05, thus the normality assumption is also violated.



Graph2.1.1 Check assumptions of 5th order polynomial model

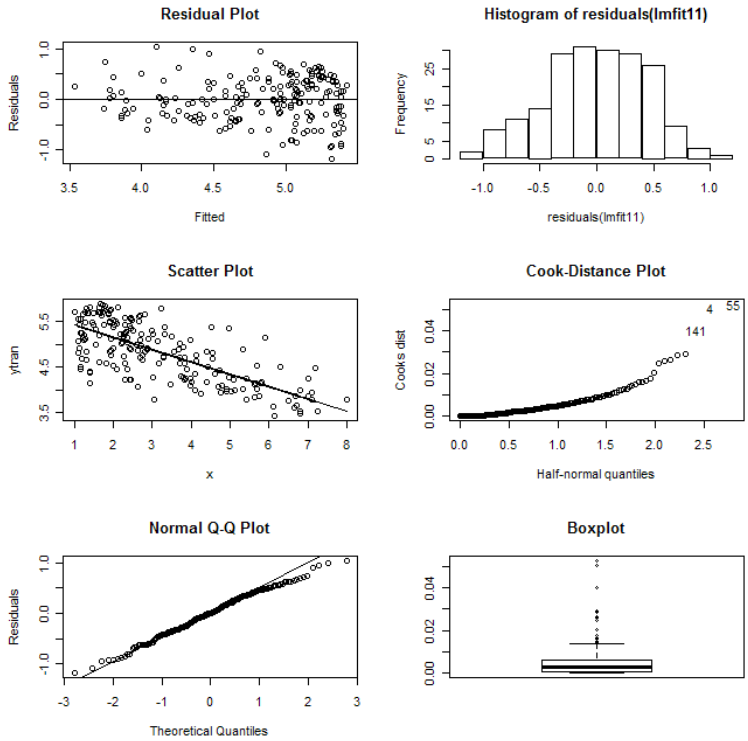
From the residual plot, I see the variance increases as fitted y increases, which implies that variance depends on x_i and thus constant-variance assumption is violated. From both of normal Q-Q plot and the histogram, I see normality is not satisfied in this model. The scatter plot implies that the 5th order polynomial model is not good enough to fit the relationship between the two variables. The half-normal plot and boxplot of Cook-distances indicate that there are some influential points and outliers in this model, respectively.

2.1.3 Box-Cox transformation

To deal with non-constant-variance and non-normality, Box-Cox transformation is one of the best methods. The estimated λ of Box-Cox transformation is -0.126, so original gene expression values y's are transformed as $\mathrm{g(y)} = y^{-0.126} - 1/(-0.126)$. The result of fitting transformed y on $\mathrm{x} + x^2 + x^3 + x^4 + x^5$ is only intercept and coefficient of x are significant. Thus, the simple linear regression of transformed y on x should be fitted. The results are $\mathrm{g(y)} = 5.69391 - 0.26998 * \mathrm{x}$, $R^2 = 0.5297, MSE = \hat{\sigma}^2 = 0.191844$. Both of the intercept and coefficient of x are significant in this case.

Next, assumptions should be checked on the transformed model. The p-values of Run-test and Durbin-Watson test are 0.9428 and 0.7761, respectively. So the randomness assumption is satisfied in this model. The p-value of Shapiro-Wilk normality test is 0.2765, thus normality assumption is also guaranteed. P-value of studentized Breusch-Pagan test is 0.05653, which is slightly larger than 0.05 and the proportion that two random selected groups of residuals with the same sample size having different variances is 0.0478, which is significantly larger than 0.

Therefore, there is not adequate evidence that constant-variance assumption is satisfied.



Graph2.1.2 Check assumptions after Box-Cox transformation

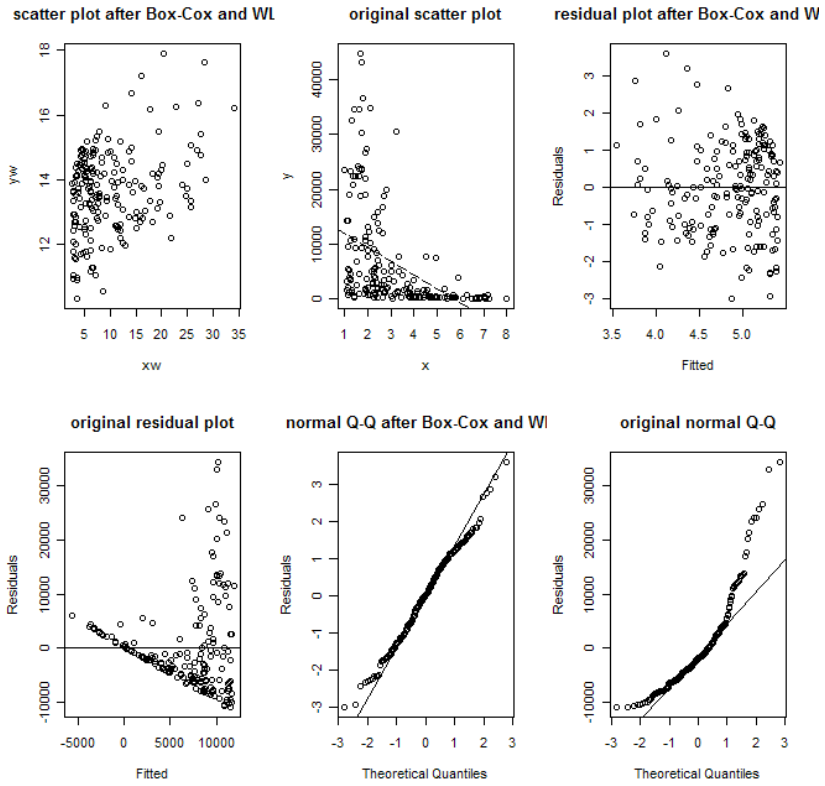
These graphs are also evidence that the linearity, normality and constant-variance property have all been improved greatly after Box-Cox transformation. But we can further improve the conformity of constant-variance assumption.

2.1.4 Weighted Least Square

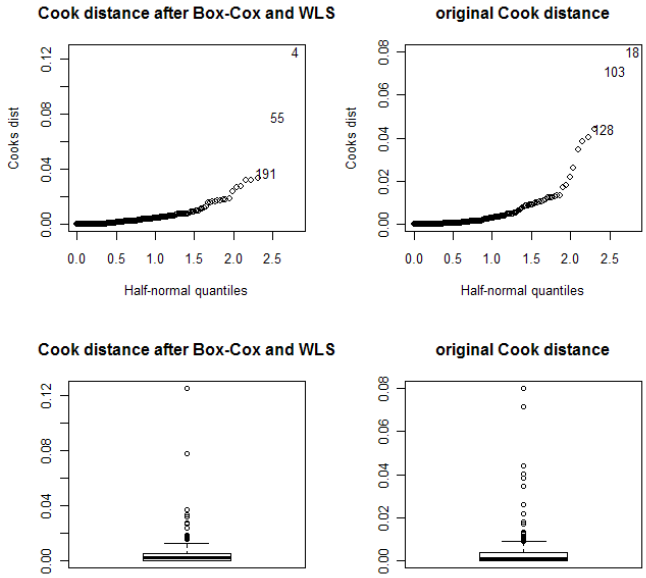
Now proper statistical methods are needed to further transform the model to satisfy the constant-variance assumption. WLS is a good choice in this situation. The fitted WLS model is $w^{1/2}$g(y)= $w^{1/2}$*(5.68768 -0.26811*x), here w is a diagonal matrix with $w_{ii}=1/(\widehat{absolute(residual)})^2$. And R^2= 0.5864, MSE=$\hat{\sigma}^2$= 1.520289.

2.1.5 Conclusions

After Box-Cox transformation and WLS, the p-value of Run-test is 0.9428, so the randomness assumption is satisfied in this model. The p-value of Shapiro-Wilk normality test is 0.3373, which is larger than the p-value before using WLS, thus normality assumption is guaranteed. The p-value of studentized Breusch-Pagan test is 0.05653 and the proportion that two random selected groups of residuals with the same sample size having different variances is 0.0471, smaller than the one before using WLS.



Graph2.1.3 Comparison of residual plots, normal Q-Q plots and scatter plots



Graph2.1.4 Comparison of cook-distance

From the comparison between scatter plots, normal Q-Q plots and residual plots of simple linear regression using original data and transformed data through Box-Cox transformation and WLS, we see that the scatter plot presents more significant linear pattern now; the variance is far more constant and normal Q-Q line is also much closer to the y=x line. Thus, one of the best

model is obtained by first fitting simple linear regression of transformed y on x, then using WLS to further fit the model.

2.1.6 Statistical inference based on fitted model

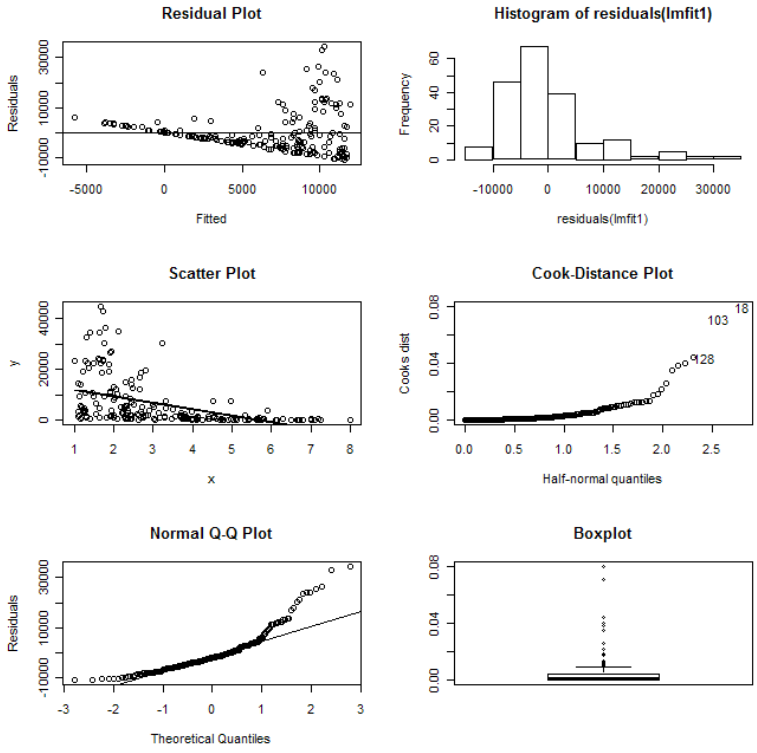
95% CI for β_0 and β_1 in the final model (after Box-Cox transformation and WLS) are (5.5546117, 5.8207539) and (-0.3002495, -0.2359805), respectively. Given x_{new}=3, 95% predict interval for y_{new} is (3.699853×10^{-101}, $+\infty$).

2.2 Fit the best model with outliers removed

2.2.1 Fit simple linear model

This time, I first fit the simple linear model to observe the outliers in 1st order model. The fitted simple linear model is $\hat{y} = 14478.6 - 2530.3 * x$. And R^2= 0.2147, MSE=σ̂^2= 69422224.

2.2.2 Removing the 3 most significant outliers



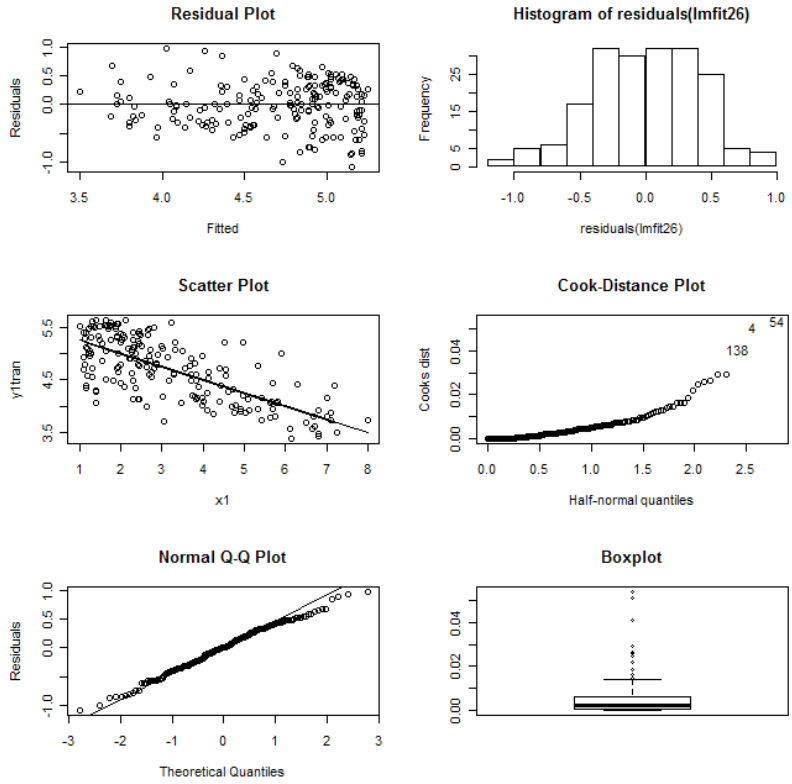
Graph 2.2.1 Check assumptions of simple linear model

From the residual plot, I see the variance increases as fitted y increases, which implies that variance depends on x_i and thus constant-variance assumption is violated. From both of normal Q-Q plot and the histogram, I see normality is not satisfied in this model. The scatter plot implies that the simple linear model is not a good fit of the relationship between the two variables. The half-normal plot and boxplot of Cook-distances indicate that there are some influential points and outliers in this model, respectively, especially the 18th, 103~~th~~ and 128th observations. Thus

they will be discarded from further analysis. Next, I fit a quadratic model $\hat{y}$ =18721.8 -5947.5*x+470.4*x^2 without the 3 outliers and obtain R^2= 0.2501, MSE=σ̂^2= 52635025.
The p-values of run-test and Durbin-Watson test are 0.5606 and 0.3639, respectively, both of which are larger than 0.05, so the randomness assumption can be accepted at the significance lever 0.05.
The p-value of studentized Breusch-Pagan test is 3.905e-07, which is significantly smaller than 0.05 and the proportion that two random selected groups of residuals with the same sample size having different variances is 0.0478, which is significantly larger than 0, thus, we can reach the conclusion that constant-variance assumption is violated.
The p-value of Shapiro-Wilk normality test is 2.089e-11, which is far less than 0.05, thus the normality assumption is also violated.

2.2.3 Box-Cox transformation after removing outliers.

Then I use Box-Cox transformation to deal with non-constant-variance and non-normality. The estimated λ of Box-Cox transformation is -0.134, so original gene expression values y's are transformed as g(y)=(y^(-0.134)-1)/((-0.134)). The result of fitting transformed y on x+x^2 is that only intercept and coefficient of x are significant. Thus, the simple linear regression of transformed y on x should be fitted. The results are g(y)= 5.49830-0.25030 *x, R^2= 0.5278, MSE=$\hat{\sigma}^2$= 0.166954. Both of the intercept and coefficient of x are significant in this case.
Next, assumptions should be checked on the transformed model. The p-values of Run-test and Durbin-Watson test are 0.6625 and 0.7879, respectively. So the randomness assumption is satisfied in this model. The p-value of Shapiro-Wilk normality test is 0.2878, p-value of studentized Breusch-Pagan test is 0.09547, both of which are slightly larger than those of the best method proposed in section 2.1; the proportion that two random selected groups of residuals with the same sample size having different variances is 0.0494. Therefore both of normality and constant-variance assumptions can be accepted.



Graph 2.2.2 Check assumptions after removing 3 outliers and Box-Cox transformation

From the residual plot, I see this model is much more close to constant-variance after removing three most significant outliers and Box-Cox transformation. The normal Q-Q plot and histogram indicate that the normality has been improved greatly.

The best model in this section is obtained by first deleting three most significant outliers in the simple linear regression on original data, then using Box-Cox transformation to further decreasing non-constant variance and non-normality.

2.2.4 Conclusions

Based on the above tests and graphs, removing outliers from the original simple linear regression, then fitting simple linear regression of transformed y on x(because in this case only the 1st order model is significant between g(y) and x) can lead to one of the best models in fitting the relationship between biomarker values and gene expression values.

2.2.5 Statistical Inference

95% CI for β_0 and β_1 in the proposed linear regression after removing 3 outliers and Box-Cox transformation of y are (5.3742276, 5.6223656), (-0.2843664 ,-0.2162398), respectively. Given $x_{new} = 2$, the 95% predict interval for y_{new} is (-3.321433$\times 10^{13}$, 3.321433$\times 10^{13}$).

3.Appendix(R code)

#import R packages and data

```
library(alr3)
library(MASS)
library(zoo)
library(lmtest)
library(lattice)
library(grid)
library(faraway)
data<-read.table("I:/Statistics/stat5044 takehome/24tpdata.txt",header=T)
x<-data[,1]
y<-data[,2]
```

#descriptive statistics of variables x and y

```
summary(x)
sd(x)
summary(y)
sd(y)
par(mfrow=c(2,2))
boxplot(x,ylab='biomarker value')
boxplot(y,ylab='gene expression value')
hist(x,freq=FALSE,xlab='biomarker value')
hist(y,freq=FALSE,xlab='gene expression value')
```

simple linear regression and scatter plot between x and y

```
lmfit1<-lm(y~x)
summary(lmfit1)
par(mfcol=c(1,1))
plot(x,y)
abline(coef(lmfit1),lty=5)
```

#fit polynomial regression models between x and y

```
lmfit2<-lm(y~x+I(x^2))
summary(lmfit2)
lmfit3<-lm(y~x+I(x^2)+I(x^3))
summary(lmfit3)
lmfit4<-lm(y~x+I(x^2)+I(x^3)+I(x^4))
summary(lmfit4)
lmfit5<-lm(y~x+I(x^2)+I(x^3)+I(x^4)+I(x^5))
summary(lmfit5)
lmfit6<-lm(y~x+I(x^2)+I(x^3)+I(x^4)+I(x^5)+I(x^6))
summary(lmfit6)
lmfit7<-lm(y~x+I(x^2)+I(x^3)+I(x^4)+I(x^5)+I(x^6)+I(x^7))
summary(lmfit7)
lmfit8<-lm(y~x+I(x^2)+I(x^3)+I(x^4)+I(x^5)+I(x^6)+I(x^7)+I(x^8))
```

```
summary(lmfit8)
lmfit9<-lm(y~x+I(x^2)+I(x^3)+I(x^4)+I(x^5)+I(x^6)+I(x^7)+I(x^8)+I(x^9))
summary(lmfit9)
lmfit10<-lm(y~x+I(x^2)+I(x^3)+I(x^4)+I(x^5)+I(x^6)+I(x^7)+I(x^8)+I(x^9)+I(x^10))
summary(lmfit10)
lmfit11<-lm(y~x+I(x^2)+I(x^3)+I(x^4)+I(x^5)+I(x^6)+I(x^7)+I(x^8)+I(x^9)+I(x^10)+I(x^11))
summary(lmfit11)
lmfit12<-lm(y~x+I(x^2)+I(x^3)+I(x^4)+I(x^5)+I(x^6)+I(x^7)+I(x^8)+I(x^9)+I(x^10)+I(x^11)+I(x^12))
summary(lmfit12)
```

#check assumptions of 11th order polynomial regression using graphs

```
par(mfcol=c(2,3))
plot(fitted(lmfit11),residuals(lmfit11),xlab="Fitted",ylab="Residuals")
abline(h=0)
plot(fitted(lmfit11),abs(residuals(lmfit11)),xlab="Fitted",ylab="|Residuals|" )
qqnorm(residuals(lmfit11),ylab="Residuals")
qqline(residuals(lmfit11))
hist(residuals(lmfit11))
cook<-cooks.distance(lmfit11)
halfnorm(cook,3,ylab="Cooks dist")
boxplot(cook)
```

#check assumptions of 11th order polynomial regression using tests

```
runs.test(residuals(lmfit11))
dwtest(lmfit11)
bptest(lmfit11)
shapiro.test(residuals(lmfit11))
```

#check assumptions of 5^{th} order polynomial regression using graphs

```
postscript("I:/Statistics/stat5044 takehome/check5.ps")
par(mfcol=c(2,3))
plot(fitted(lmfit5),residuals(lmfit5),xlab="Fitted",ylab="Residuals",main="Residual Plot")
abline(h=0)
plot(x,y,main="Scatter Plot")
f<-function(x){-56858.43+117028.72*x-68778.47*x^2+17423.50*x^3-2008.38*x^4+86.42*x^5}
y5<-f(x)
lines(x,y5)
qqnorm(residuals(lmfit5),ylab="Residuals")
qqline(residuals(lmfit5))
hist(residuals(lmfit5))
cook<-cooks.distance(lmfit5)
halfnorm(cook,3,ylab="Cooks dist",main="Cook-Distance Plot")
boxplot(cook,main="Boxplot")
dev.off()
```

#check assumptions of 5^{th} order polynomial regression using tests

```
runs.test(residuals(lmfit5))
dwtest(lmfit5)
bptest(lmfit5)
shapiro.test(residuals(lmfit5))

pv<-rep(0,10000)
 for(i in 1:10000){sm<-sample(1:193,192,replace=FALSE)
 sm1<-sm[1:96]
 sm2<-sm[97:192]
 g1<-residuals(lmfit5)[sm1]
 g2<-residuals(lmfit5)[sm2]
 d1<-abs(g1-median(g1))
 d2<-abs(g2-median(g2))
 pv[i]<-t.test(d1,d2)$p.value}
nonconstantvar<-sum(pv<0.05)
nonconstantvar
```

estimate λ in Box-Cox transformation for the 5^{th} order polynomial regression

```
par(mfrow=c(1,1))
boxcox(lmfit5,plotit=T)
boxcox(lmfit5,plotit=T,lambda=seq(-0.13,-0.12,by=0.001))
```

#fit transformed 5^{th} order polynomial regression

```
lam<--0.126
ytran<-(y^lam-1)/lam
lmfit51<-lm(ytran~x+I(x^2)+I(x^3)+I(x^4)+I(x^5))
summary(lmfit51)
```

#fit transformed simple linear regression

```
lmfit11<-lm(ytran~x)
summary(lmfit11)
```

#check assumptions of simple linear regression of transformed model using tests

```
runs.test(residuals(lmfit11))
dwtest(lmfit11)
bptest(lmfit11)
shapiro.test(residuals(lmfit11))
```

#check assumptions of simple linear regression of transformed model using graphs

```
par(mfcol=c(3,2))
 plot(fitted(lmfit11),residuals(lmfit11),xlab="Fitted",ylab="Residuals",main="Residual Plot")
abline(h=0)
```

```
plot(x,ytran,main="Scatter Plot")
f<-function(x){5.69391-0.26998*x}
y23<-f(x)
lines(x,y23)
qqnorm(residuals(lmfit11),ylab="Residuals")
qqline(residuals(lmfit11))
hist(residuals(lmfit11))
cook<-cooks.distance(lmfit11)
halfnorm(cook,3,ylab="Cooks dist",main="Cook-Distance Plot")
boxplot(cook,main="Boxplot")

#preparation for estimating w
resid11<-residuals(lmfit11)
absresid11<-abs(resid11)
plot(x,absresid11)

#estimation of w and fit WLS model
lmfitw0<-lm(absresid11~x)
w<-1/(fitted(lmfitw0))^2
wlm11<-lm(ytran~x,weights=w)
summary(wlm11)
confint(wlm11)

# comparisons between graphs before and after Box-Cox, WLS technique
postscript("I:/Statistics/stat5044 takehome/comparewlm11.ps")
par(mfrow=c(2,3))
yw<-w^0.5*ytran
xw<-w^0.5*x
wresid<-w^0.5*residuals(lmfit11)
plot(xw,yw,main='scatter plot after Box-Cox and WLS')
abline(coef(wlm11),lty=10)
plot(x,y,main='original scatter plot')
abline(coef(lmfit1),lty=5)
plot(fitted(wlm11),wresid,xlab="Fitted",ylab="Residuals",main='residual plot after Box-Cox and
WLS')
abline(h=0)
plot(fitted(lmfit1),residuals(lmfit1),xlab="Fitted",ylab="Residuals",main="original residual plot")
abline(h=0)
qqnorm(wresid,ylab="Residuals",main="normal Q-Q after Box-Cox and WLS")
qqline(wresid)
qqnorm(residuals(lmfit1),ylab="Residuals",main="original normal Q-Q")
qqline(residuals(lmfit1))

#Check assumptions of the model after Box-Cox transformation and WLS
```

```
runs.test(wresid)
bptest(wlm11)
shapiro.test(wresid)

pv<-rep(0,10000)
for(i in 1:10000){sm<-sample(1:193,192,replace=FALSE)
sm1<-sm[1:96]
sm2<-sm[97:192]
g1<-wresid[sm1]
g2<-wresid[sm2]
d1<-abs(g1-median(g1))
d2<-abs(g2-median(g2))
pv[i]<-t.test(d1,d2)$p.value}
nonconstantvar<-sum(pv<0.05)
nonconstantvar
```

#comparisons between Cook-distance before and after Box-Cox, WLS technique

```
par(mfcol=c(2,2))
cook<-cooks.distance(wlm11)
halfnorm(cook,3,ylab="Cooks dist",main=" Cook distance after Box-Cox and WLS")
boxplot(cook,main="Cook distance after Box-Cox and WLS")
cook<-cooks.distance(lmfit1)
halfnorm(cook,3,ylab="Cooks dist",main="original Cook distance")
boxplot(cook,main="original Cook distance")
```

CI for β_0 and β_1 in the final model(after Box-Cox transformation and WLS)

```
confint(wlm11)
```

PI for y_{new} given x_{new}=3

```
xnew=3
sxx=sum(x-mean(x))^2
gyh=5.68768-0.26811*xnew
a1<-gyh-qt(0.975,191)*sqrt(1+1/193+(xnew-mean(x))^2/sxx)
a2<-gyh+qt(0.975,191)*sqrt(1+1/193+(xnew-mean(x))^2/sxx)
c1<-(-0.126*a1+1)^(1/-0.126)
c2<-(-0.126*a2+1)^(1/-0.126)
ci<-c(c1,c2)
ci
```

#fit original simple linear model

```
lmfit1<-lm(y~x)
summary(lmfit1)
```

#check assumptions of original simple linear regression using graphs

```
par(mfcol=c(3,2))
plot(fitted(lmfit1),residuals(lmfit1),xlab="Fitted",ylab="Residuals",main="Residual Plot")
abline(h=0)
plot(x,y,main="Scatter Plot")
f<-function(x){14478.6-2530.3*x}
y23<-f(x)
lines(x,y23)
qqnorm(residuals(lmfit1),ylab="Residuals")
qqline(residuals(lmfit1))
hist(residuals(lmfit1))
cook<-cooks.distance(lmfit1)
halfnorm(cook,3,ylab="Cooks dist",main="Cook-Distance Plot")
boxplot(cook,main="Boxplot")
```

#fit quadratic regression without the 3 most significant outliers

```
data1<-read.table("I:/Statistics/stat5044 takehome/24tpdata1.txt",header=T)
x1<-data1[,1]
y1<-data1[,2]
lmfit24<-lm(y1~x1)
summary(lmfit24)
```

#Box-Cox transformation of the above quadratic regression

```
par(mfrow=c(1,1))
boxcox(lmfit24,plotit=T)
boxcox(lmfit24,plotit=T,lambda=seq(-0.135,-0.13,by=0.001))

lam<--0.134
y1tran<-(y1^lam-1)/lam
lmfit25<-lm(y1tran~x1+I(x1^2))
summary(lmfit25)
lmfit26<-lm(y1tran~x1)
summary(lmfit26)
```

#Check assumptions of transformed model without outliers using graphs

```
par(mfcol=c(3,2))
plot(fitted(lmfit26),residuals(lmfit26),xlab="Fitted",ylab="Residuals",main="Residual Plot")
abline(h=0)
plot(x1,y1tran,main="Scatter Plot")
f<-function(x){5.49830-0.25030*x}
y23<-f(x)
lines(x,y23)
qqnorm(residuals(lmfit26),ylab="Residuals")
qqline(residuals(lmfit26))
hist(residuals(lmfit26))
```

```
cook<-cooks.distance(lmfit26)
halfnorm(cook,3,ylab="Cooks dist",main="Cook-Distance Plot")
boxplot(cook,main="Boxplot")
```

#Check assumptions of transformed model without outliers using tests

```
runs.test(residuals(lmfit26))
dwtest(lmfit26)
bptest(lmfit26)
shapiro.test(residuals(lmfit26))

pv<-rep(0,10000)
  for(i in 1:10000){sm<-sample(1:190,190,replace=FALSE)
  sm1<-sm[1:95]
  sm2<-sm[96:190]
  g1<-residuals(lmfit26)[sm1]
  g2<-residuals(lmfit26)[sm2]
  d1<-abs(g1-median(g1))
  d2<-abs(g2-median(g2))
  pv[i]<-t.test(d1,d2)$p.value}
nonconstantvar<-sum(pv<0.05)
nonconstantvar
```

CI for $\beta_0 \; and \; \beta_1$ in the transformed simple linear model without the 3 outliers

```
confint(lmfit26)
```

95% CI for y_{new} given x_{new}=2

```
beta0h<-5.49830
beta1h<--0.25030
xnew<-2
sxx<-sum(x1-mean(x1))^2
a1<-beta0h+beta1h*xnew-qt(0.975,188)*0.4086*sqrt(1+1/190+(xnew-mean(x))^2/sxx)
a2<-beta0h+beta1h*xnew+qt(0.975,188)*0.4086*sqrt(1+1/190+(xnew-mean(x))^2/sxx)
c1<-(-0.134*a1+1)^1/(-0.134)
c2<-(-0.134*a2+1)^1/(-0.134)
CI<-c(c1,c2)
CI
```